The Annals of Statistics
2021, Vol. 49, No. 4, 2271–2291
https://doi.org/10.1214/20-AOS2036

MONITORING FOR A CHANGE POINT IN A SEQUENCE OF DISTRIBUTIONS

BY LAJOS HORVÁTH[1], PIOTR KOKOSZKA[2] AND SHIXUAN WANG[3]

[1]*Department of Mathematics, University of Utah, horvath@math.utah.edu*
[2]*Department of Statistics, Colorado State University, piotr.kokoszka@colostate.edu*
[3]*Department of Economics, University or Reading, shixuan.wang@reading.ac.uk*

We propose a method for the detection of a change point in a sequence $\{F_i\}$ of distributions, which are available through a large number of observations at each $i \geq 1$. Under the null hypothesis, the distributions F_i are equal. Under the alternative hypothesis, there is a change point $i^* > 1$, such that $F_i = G$ for $i \geq i^*$ and some unknown distribution G, which is not equal to F_1. The change point, if it exists, is unknown, and the distributions before and after the potential change point are unknown. The decision about the existence of a change point is made sequentially, as new data arrive. At each time i, the count of observations, N, can increase to infinity. The detection procedure is based on a weighted version of the Wasserstein distance. Its asymptotic and finite sample validity is established. Its performance is illustrated by an application to returns on stocks in the S&P 500 index.

1. Introduction. We propose a method for sequential detection of a change point in a sequence of distributions. Such sequences occur in many applications. Plentiful examples arise in economics and finance, including income distributions and various return distributions. We have been motivated by cross-sectional market returns, which form perhaps the most extensively studied sequence of distributions. Mathematically, we are dealing with sequences $\{F_i\}$, $\{Q_i\}$ and $\{f_i\}$ of respectively cumulative distribution, quantile or density functions, each offering an equivalent data model. The problem we consider is as follows. Under the null hypothesis, at each time period $i \geq 1$, the distributions F_i are equal. Under the alternative hypothesis, there is a change point $i^* > 1$, such that $F_i = G$ for $i \geq i^*$ and some unknown distribution G, which is not equal to F_1. The change point, if it exists, is unknown, and the distributions before and after the potential change point are unknown. The decision about the existence of a change point must be done sequentially, as new data become available. Precise formulation is presented in Section 2.

The problem we consider is different from the well-studied problem of sequential detection of a change in distribution based on only a single scalar observation at each time period i. This problem was studied, in Pollak (1985), Yakir (1997) and Polunchenko and Tartakovsky (2010), among others. Their work focuses on mini-max optimality. In our setting, we have a very large number of observations at every time i, rather than just one. Our theory includes asymptotic analysis as the number of observations in each period tends to infinity. It thus falls to the broad domain of the analysis of high-dimensional time series.

In applications, none of the mathematical models, be it the distribution or the quantile function, is directly accessible. The data available at each time instant i are scalar observations $X_{i,j}$. We focus on the detection of change in the quantile function. The justification for using the quantile function is discussed in the following. The large sample theory, as the

Received April 2020; revised November 2020.
MSC2020 subject classifications. Primary 62G30, 62L10; secondary 62G10, 62G20.
Key words and phrases. Change point detection, Empirical quantile function, Sequential monitoring, Wasserstein distance.

number of observations $X_{i,j}$ in each period i tends to infinity, is quite complex. It connects to profound results on the asymptotic behavior of the *empirical* quantile function, which were established over the last three decades.

Before discussing our theory in greater detail, we note that change point analysis has seen renewed interest over the last decade, mainly due to the need to create change point detection tools suitable for complex, non-Euclidean data structures and high-dimensional and functional data. Discussing, or even listing, dozens of relevant contributions is not possible, so we merely list a few, admittedly subjectively selected, recent contributions which contain further references. Bardsley et al. (2017) and Gromenko, Kokoszka and Reimherr (2017) offer different perspectives on change point analysis of functional data. Jirak (2015), Barigozzi, Cho and Fryzlewicz (2018) and Chen, Wang and Samworth (2020) study change point detection in high-dimensional time series. Dubey and Müller (2020) propose a method of change point detection suitable for general metric spaces and provide a good discussion of change point detection for complex data structures, including networks. Special challenges arise when multiple change points might be present, Li and Jin (2018) discuss them and propose a novel solution. These papers, with the exception of Chen, Wang and Samworth (2020), deal with detecting change points in a historical sample. Sequential change point detection in complex structures has attracted a great deal of attention in engineering literature, focusing on practical aspects, but profound mathematical theory also exists; see, for example, Xie and Siegmund (2013). The recent work of Padilla et al. (2019) is closely related. It is concerned with detecting a change point if a sequence of densities, $\{f_i\}$, assuming each density is available through a fine histogram obtained from a very large and unknown number of observations $X_{i,j}$. Their approach is to find a method with a minimal detection delay under a constraint on the rate of false alarms in a specified time window. Mathematically, the constraint is imposed on the expected number of false alarms. In our approach, we follow the paradigm of Chu, Stinchcombe and White (1996), introduced in the context of economic data, and impose a constraint on the probability of a false alarm, which has essentially the same interpretation as the size of a Neyman–Pearson test. Our theoretical and empirical analysis includes the effect of the estimation in the training sample. Because of this and the emphasis on type I error, our approach is not directly comparable to that of Padilla et al. (2019). Since no parametric assumptions are imposed, our approach is also completely different from methods based on likelihood ratios whose theoretical justification is based on mini-max optimality.

The main advance over existing work that uses the paradigm of Chu, Stinchcombe and White (1996) is that we handle monitoring for a change point in a sequence of distributions available indirectly trough a large number of observations at every time point. Previous work has considered only sequences of scalars arising from various models. To accommodate this more complex setting, we must choose a function describing the distribution and based on it construct a suitable detector. Theory needed to provide an asymptotic justification is much more complex than in the scalar case. We conclude this section with highlighting some of these points.

Each mathematical model for the distribution of scalar observations, that is, the cumulative distribution, quantile and density function, is subject to well-known constraints. Of these three objects, the quantile function is subject to least constraints. It enters directly into the definition of the Wasserstein distance,

$$d^2(G,F)=\int_0^1 \bigl(G^{-1}(t)-F^{-1}(t)\bigr)^2\,dt,$$

between two cumulative distribution functions G and F. The above distance is also known as the Kantorovich–Wasserstein or L_2-Wasserstein distance. Panaretos and Zemel (2019) provide a survey of its applications to statistics and data science. It is used in the context

of functional data analysis of densities in Petersen and Müller (2016) and in a broad setting in Panaretos and Zemel (2016), to name relatively recent contributions. We work with a more general version of this distance. The Lebesgue measure dt is replaced by $w(t)\,dt$. The weight function is needed to accommodate various types of tail behavior of the observations $X_{i,j}$, as the count of observations indexed by j increases. In specific applications, various weight functions can be used to emphasize either the center of the distributions or the extreme tails.

Suppressing the index i, denote by F_N the empirical distribution function of X_j, $1 \le j \le N$, and by Q_N the corresponding quantile function. To establish asymptotic validity of our monitoring procedure, we must study the asymptotic behavior of the first three moments of

$$d_w^2(F_n, F) = \int_{1/(2N)}^{1-1/(2N)} \big(Q_N(t) - Q(t)\big)^2 w(t)\,dt,$$

where Q is the population quantile function. The reason for it is explained in Section 2, the corresponding results are presented in Supplementary Material (Horváth, Kokoszka, and Wang (2021)). There is nontrivial interplay between the behavior of Q as $t \to 0$ and $t \to 1$ and the behavior of the weight function w at these end points. Related results are proven in del Barrio et al. (1999), del Barrio, Giné and Matrán (1999) and del Barrio, Giné and Utzet (2005). They pertain to weak convergence rather than the convergence of moments. For example, del Barrio et al. (1999) establish limiting behavior of $R_N = S_N^{-2} d^2(F_n, \mathcal{H})$, where $\mathcal{H}$ is the subspace spanned by the normal distributions, and S_N is the sample variance of the X_j. Their results are motivated by normality tests of the Shapiro–Wilk type, so they assume that the X_j are normal. In that case, $N(R_N - a_N)$ converges to a functional of the Brownian bridge, $B(t)$, $0 \le t \le 1$; see their Theorem 2. The centering constants have the form

$$a_N = \frac{1}{N} \int_{1/(N+1)}^{N/(N+1)} \frac{t(1-t)}{\phi^2(\Phi^{-1}(t))}\,dt.$$

We obtain convergence of moments for general classes of distributions. For example, for the first moment we obtain

$$N E d_w^2(F_n, F) \to \int_0^1 \frac{t(1-t)}{f^2(Q(t))} w(t)\,dt$$

for weight functions w matching the behavior of the quantile function Q at the end points of the interval $[0, 1]$. In the above formulas, ϕ is the standard normal density and f is the common density of the X_j in our setting. In a recent paper, Berthet and Fort (2020) obtained almost sure laws for the Wasserstein metric assuming normal observations. Csörgő and Horváth (1993) provide several results for L_p norms of the difference between empirical and theoretical quantiles using the weight functions $w(t) = 1$ and $w(t) = 1/Q'(t)$. It is hoped that our fairly general results (basically an arbitrary quantile function Q) might prove useful in theoretical work relying on the convergence of general Wasserstein distances between empirical and population distributions.

The remainder of the paper is organized as follows. In Section 2, we rigorously formulate the monitoring (sequential detection) problem, describe our procedure and state results establishing its asymptotic validity. Section 3 is dedicated to an empirical study of cross-sectional returns of the constituent stocks in the S&P 500 index. It illustrates the practical usefulness of our method. Finite sample performance is investigated in Section 4. The Supplementary Material contains the proofs of the results stated in Section 2.

2. Assumptions and main results. We assume that at time i we have N observations $X_{i,1}, X_{i,2}, \ldots, X_{i,N}$. Throughout this paper, we assume that the following assumption holds.

ASSUMPTION 2.1. The random variables $X_{i,j}$, $1 \le i < \infty$, $1 \le j \le N$ are independent.

We wish to test the null hypothesis:

$$H_0 : X_{i,j}, 1 \le i < \infty, 1 \le j \le N \text{ are identically distributed.}$$

We assume that over the initial M time points the distribution is constant. This is formalized in the following assumption.

ASSUMPTION 2.2. There is $M \ge 1$ such that the $X_{i,j}$, $1 \le i \le M$, $1 \le j \le N$ are identically distributed.

Assumption 2.2 means that H_0 always holds for the training sample. Under the alternative the distribution of the observations changes at time $M + k^*$, where k^* is the unknown time of change:

$$\begin{aligned} H_A :& X_{i,j}, 1 \le i < M + k^*, 1 \le j \le N \text{ are identically distributed,} \\ & X_{i,j}, M + k^* \le i < \infty, 1 \le j \le N \text{ are identically distributed} \\ & \text{but the distribution of } X_{1,1} \text{ differs from that of } X_{M+k^*,1}. \end{aligned}$$

If $k^* = 1$, then the distribution of the observations changes immediately after the training sample. In our model, at time i, we have an independent and identically distributed random sample. Under the null hypothesis, the common distribution of the observations remains the same. Under the alternative hypothesis, at an unknown time $M + k^*$ the distribution of the sample changes to a different one.

Our detection method is based on the quantiles (order statistics). For any $i \ge 1$, let

$$X_{i;1,N} \le X_{i;2,N} \le \cdots \le X_{i;N,N}$$

denote the order statistics of $X_{i,j}$, $1 \le j \le N$. Following Csörgő and Révész (1981) (cf. also Csörgő and Horváth (1993)), we define the quantile function at time ith as

$$Q_{i,N}(t) = X_{i;j,N} \quad \text{if } (j-1)/N \le t < j/N, j = 1, 2, \ldots, N.$$

We will use the average quantile function of the training sample defined by

$$\bar{Q}_M(t) = \frac{1}{M} \sum_{i=1}^{M} Q_{i,N}(t), \quad 0 \le t < 1.$$

We measure the deviation from the average quantile function of the training sample by

$$\xi_{i,N} = \int_{1/(2N)}^{1-1/(2N)} \{Q_{i,N}(t) - \bar{Q}_M(t)\}^2 w(t)\, dt, \quad 1 \le i < \infty, \tag{1}$$

where $w \ge 0$ is a weight function. If $w(t) = 1$, then $\sqrt{\xi_{i,N}}$ is the Kantorovich–Wasserstein or minimal L_2 distance between the measures associated with $Q_{i,N}(t)$ and $\bar{Q}_M(t)$.

We compute the mean and the variance of the discrepancy measures in the training sample:

$$\hat{\xi}_{M,N} = \frac{1}{M} \sum_{i=1}^{M} \xi_{i,N} \quad \text{and} \quad \hat{\sigma}^2_{M,N} = \frac{1}{M-1} \sum_{i=1}^{M} (\xi_{i,N} - \hat{\xi}_{M,N})^2.$$

Our sequential method is based on a detector and a boundary function within a general paradigm proposed by Chu, Stinchcombe and White (1996). The detector is given by

$$\Gamma(M,s)=\frac{1}{\hat{\sigma}_{M,N}}\left|\sum_{i=M+1}^{M+s}\xi_{i,N}-\frac{s}{M}\sum_{i=1}^{M}\xi_{i,N}\right|,\quad 1\le s<\infty. \tag{2}$$

The form of the detector is fairly intuitive. We compare the average of the weighted distances (1) in the monitoring sample to their average in the training sample. The absolute difference is normalized to ensure a variance-free limit. Only specific boundary functions will ensure that the probability of a false rejection can be controlled. We use the boundary function

$$g(M,s)=cM^{1/2}\left(1+\frac{s}{M}\right)\left(\frac{s}{M+s}\right)^{\gamma}, \tag{3}$$

where γ is a constant satisfying

ASSUMPTION 2.3. $0\le\gamma<1/2$.

The form of the boundary function is motivated by the objective that the limit distribution has a simple form and can be easily simulated. It would be possible to replace g with

$$g(M,s)=cM^{1/2}\left(1+\frac{s}{M}\right)\left(\frac{s}{M+s}\right)^{\gamma}\ell\left(\frac{s}{M+s}\right),$$

where is ℓ is a slowly varying at ∞. Including the function ℓ is simple from a theoretical point of view if $0\le\gamma<1/2$ and harder if $\gamma=1/2$. From the practical point of view, the slowly varying function does not enhance the statistical procedure; in most applications, slowly varying functions are set to be constants. We will see that our procedure is asymptotically valid for any γ satisfying Assumption 2.3. In finite samples, we have the freedom to choose it to tune the detection procedure to ensure accurate calibration under H_0 for small and moderate M and increase the power under the alternative.

A change point in distribution is detected if the detector crosses the boundary function. The stopping time is thus defined as follows:

$$\tau_M=\begin{cases}\inf\{s\ge 1:\Gamma(M,s)>g(M,s)\},\\ \infty\quad \text{if } \Gamma(M,s)\le g(M,s) \text{ for all } s\ge 1.\end{cases}$$

In our sequential detection approach, the key is to control the rate of false alarms. This is expressed as the following asymptotic condition:

$$\lim_{M\to\infty}P\{\tau_M<\infty\}=\alpha\quad\text{under } H_0, \tag{4}$$

where $0<\alpha<1$ is a prescribed significance level. According to (4), the probability of false stopping (rejection of the null hypothesis when it is correct) is α, if M is large. We will show that for any value of γ, condition (4) is met by suitably choosing $c=c(\gamma,\alpha)$ in (3). The computation of $c(\gamma,\alpha)$ is possible by expressing the limit in (4) in terms of a crossing probability for a functional of a Wiener process, specifically by showing that

$$\lim_{M\to\infty}P\{\tau_M<\infty\}=P\left\{\sup_{0<u\le 1}\frac{|W(u)|}{u^{\gamma}}\le c\right\}, \tag{5}$$

where $\{W(u),u\ge 0\}$ denotes a Wiener process (standard Brownian motion). Verification of (5) is the chief theoretical contribution of this paper. It requires suitable approximations of the partial sums of the $\xi_{i,N}$, the weighted Kantorovich–Wasserstein distances. If N is finite, these are independent and identically distributed random variables, so the approximations in

Komlós, Major and Tusnády (1975, 1976) can be used. However, if $N \to \infty$, these results cannot be used, since we need to deal with an array of random variables. In this case, we use the Skorokhod (1965) representation theorem. Hence we need to establish lower bounds for the variances and upper bounds for the third absolute moments for the L_2 weighted distances between the empirical and theoretical quantiles. These results are established in the course of the proofs of the remaining theorems of this section. We also need to prove that using $\bar{Q}_M(t)$ instead of $Q(t)$ does not change the limit results. We aimed at formulating our results in such a way that there is no connection between the length of the training period, M, and the count of observations, N, which is allowed to tend to infinity in an arbitrary manner, including nonmonotonic increase. This is motivated by the financial data we consider; if N is the count of assets, it is in no way related to the length of the training period M. Another advantage of our theory is that it covers most scenarios for the form of distribution of the observations $X_{i,j}$ (indexed by j) in a unified way. We essentially classify these scenarios by membership in a domain of attraction of one of the types of extreme value distribution. For a different domain, different assumptions on the weight function w are needed, but the final form of limit result is the same, that is, (5). The assumptions of the weight function w are weak, and are satisfied in all cases by reasonable functions w. This justifies the application of our theory without knowing much about the distribution of the $X_{i,j}$.

We begin with the case when N is a fixed number. To prove (4), we need $E\xi_{i,N}^2 < \infty$, which requires that the observations have at least four moments. Assumption 2.4 below is thus close to optimal because one needs to apply some form of a central limit theorem to the $\xi_{i,N}$, so they should have finite variance.

ASSUMPTION 2.4. $E|X_{i,j}|^{4+\delta} < \infty$ with some $\delta > 0$.

In the case of a fixed N, the requirement on weight function W are very general. It must be bounded on $[0, 1]$, positive on $(0, 1)$ and can be zero at the end points, as stated in the following assumption.

ASSUMPTION 2.5. $\inf_{\epsilon \le t \le 1-\epsilon} w(t) > 0$ for all $0 < \epsilon < 1/2$, $w(0) \ge 0$, $w(1) \ge 0$ and $\sup_{0<t<1} w(t) < \infty$.

THEOREM 2.1. *If H_0 and Assumptions 2.1–2.5 are satisfied and N is a fixed number, then relation (5) holds.*

Next, we consider cases when $N \to \infty$. For the sake of simplicity of the presentation, we assume that both tails of Q satisfy similar conditions. This is not required, (5) remains valid if the two tails behave differently, with separate sets of assumptions on each tail, and with matching assumptions on the weight function.

We begin with the case of regularly varying quantile functions. Recall that g is a regularly varying function at 0 with index α, if for all $\lambda > 0$,

$$\lim_{x \to 0} \frac{g(\lambda x)}{g(x)} = \lambda^{\alpha}.$$

ASSUMPTION 2.6. The functions $Q(t)$ and $Q(1 - t)$ are regularly varying at 0 with parameters $-\alpha_1$ and $-\alpha_2$, respectively, where $0 \le \alpha_1 < 1/4$ and $0 \le \alpha_2 < 1/4$.

We note that Assumption 2.6 implies Assumption 2.4. If the parameter of the regular variation is 0, then the function is called slowly varying. For definitions and properties of regularly and slowly varying functions, we refer to Bingham, Goldie and Teugels (1987).

We consider two sets of assumptions on the weight function w that match Assumption 2.6. Let

$$\frac{1}{h(t)} = \frac{w(t)}{f^2(Q(t))}, \quad 0 < t < 1,$$

where f denotes the density of F, and F is the distribution function of the observations. First, we consider the case when $h(t)$ is not large in a neighborhood of 0 and also of 1.

ASSUMPTION 2.7. For some $\beta < 2$ and some $c > 0$,

$$\frac{1}{h(t)} \le c\big(t(1-t)\big)^{-\beta}.$$

In case of a negative β, we are giving very little weight to the quantile functions in neighborhoods of 0 and 1. Assumption 2.7 holds if $1/h(t)$ is uniformly bounded from above by a constant. Any $\beta > 0$ and a suitable $c > 0$ can be used.

The next condition appeared first in Csörgő and Révész (1978).

ASSUMPTION 2.8. For some constant κ,

$$\sup_{0<t<1} t(1-t)\frac{|f'(Q(t))|}{f^2(Q(t))} \le \kappa.$$

THEOREM 2.2. *If H_0 and Assumptions 2.1–2.3 and 2.5–2.8 are satisfied*, $\min(N, M) \to \infty$, *then* (5) *holds*.

REMARK 2.1. Under Assumption 2.6, the support of the underlying distribution is the real line. The result in Theorem 2.2, including its proof remains true if $-\infty < Q(0)$ or $Q(1) < \infty$. Under these conditions, we need to replace Assumption (2.6) with the regular variation of $Q(t) - Q(0)$ and $Q(1) - Q(1-t)$ at 0 (cf. Corollary 3.3 in Csörgő and Horváth (1993), pp. 396 and 397).

Our proof show that under the conditions of Theorem 2.2 the extreme values do not play any role in the behavior of the distances $\xi_{i,N}$. We now consider the case when $\xi_{i,N}$ is determined by the smallest and largest order statistics. Since only the tails of $Q(t)$ matter, we need more information on the weight function $w(t)$. We thus formulate the following assumption, which can replace Assumptions 2.7 and 2.8 in Theorem 2.2.

ASSUMPTION 2.9. The functions $w(t)$ and $w(1-t)$ are regularly varying at 0 with indices τ_1 and τ_2.

THEOREM 2.3. *If H_0 and Assumptions 2.1–2.3, 2.5, 2.6 and 2.9 are satisfied*, $\tau_1 - 2\alpha_1 < 0$, $\tau_2 - 2\alpha_2 < 0$, $\min(N, M) \to \infty$, *then* (5) *holds*.

REMARK 2.2. Similar to Remark 2.1, the result of Theorem 2.3 remains valid if the regular variations of $Q(t) - Q(0)$ and $Q(1) - Q(1-t)$ are assumed in a neighborhood of 0, when $-\infty < Q(0)$ or $Q(1) < \infty$.

The conditions in Theorem 2.3 and Remark 2.2 cover two classes of the domain of attraction of extreme value distributions (cf. Section 8.13 in Bingham, Goldie and Teugels (1987)). Now we consider the third class of the domain of attraction of extreme value distributions.

ASSUMPTION 2.10. For all $x, y > 0$, $y \neq 1$,

$$\lim_{t\to\infty} \frac{Q(tx) - Q(t)}{Q(ty) - Q(t)} = \frac{\log x}{\log y}$$

and

$$\lim_{t\to\infty} \frac{Q((1-t)x) - Q(1-t)}{Q((1-t)y) - Q(1-t)} = \frac{\log x}{\log y}.$$

Assumption 2.10 covers the Gumbel domain of attraction. For a discussion and some equivalent forms of Assumption 2.10, we refer to Bingham, Goldie and Teugels (1987).

THEOREM 2.4. *If H_0 and Assumptions 2.1–2.5, 2.9 and 2.10 are satisfied, $\min(N, M) \to \infty$, then (5) holds.*

As noted above, Theorem 2.2 considers the case when $\xi_{i,N}$ is determined by the middle order statistics, while the extreme values dominate the limit in Theorems 2.3 and 2.4. Next, we study the "in between" case which is referred to as Darling–Erdős type result for integrals in Csörgő and Horváth (1993).

ASSUMPTION 2.11. The function $t(1-t)w(t)/f^2(Q(t))$ is regularly varying function at 0 and 1 with index -1, that is,

$$\frac{w(t)}{f^2(Q(t))} = \frac{1}{t^2 K_1(t)} \quad \text{and} \quad \frac{w(1-t)}{f^2(Q(1-t))} = \frac{1}{(1-t)^2 K_2(t)},$$

where $K_1(t)$, $K_2(t)$ are slowly varying functions at 0, $K_1(t) \to 0$, $K_2(t) \to 0$, as $t \to 0$.

THEOREM 2.5. *If H_0 and Assumptions 2.1–2.5, 2.8, 2.9 and 2.11 are satisfied, $\min(N, M) \to \infty$, then (5) holds.*

We conclude this section by explaining how the critical value c in (5) can be found and displaying a table with a selection of these critical values. To find $c = c(\gamma, \alpha)$, we followed the following steps:

1. Simulate 50,000 independent Wiener processes $W(u)$, where u is on a grid of 10,000 equally-spaced points in $[0, 1]$.
2. Obtain $\sup_{0\le u\le 1} |W(u)|/u^\gamma$ for each simulated trajectory of Wiener processes.
3. Find (numerical search) $c(\gamma, \alpha)$ such that

$$P\left\{ \sup_{0\le u\le 1} \frac{|W(u)|}{u^\gamma} > c(\gamma, \alpha) \right\} = \alpha.$$

Table 1 displays the critical values $c(\gamma, \alpha)$ for selected values of γ and typical significance levels α.

Before we move on to data analysis in Section 3 and a simulation study in Section 4, we remind the reader that a change point is signaled at the first time s such that $\Gamma(M, s)$, given by (2), exceeds $g(M, s)$ given by (3) with the critical value c from Table 1.

Our theory focuses on the behavior of our monitoring procedure under the null hypothesis because this is where advanced mathematical tools are needed. Regardless of various assumptions, the testing procedure is the same; we only require that $M \to \infty$ and N can be bounded or $N \to \infty$, and there is no assumption on N as a function of M. However, this is

TABLE 1
Critical values $c = c(\gamma, \alpha)$ in (3)

$\gamma \setminus \alpha$	1%	2.5%	5%	10%
0.00	2.7718	2.4628	2.2232	1.9541
0.15	2.8146	2.5473	2.2963	2.0293
0.25	2.8693	2.6208	2.3652	2.1113
0.35	2.9763	2.7233	2.4946	2.2494
0.45	3.2499	3.0038	2.7793	2.5463
0.49	3.5814	3.3135	3.0722	2.8295

not the case under the alternative. Let $Q^{(1)}$ and $Q^{(2)}$ be the quantile functions before and after the change. We also assume that $Q^{(1)}$ and $Q^{(2)}$ satisfy the conditions of one of Theorems 2.1–2.5. The quantiles should be different before and after the change, so we require

$$\int_0^1 \bigl(Q^{(1)}(u) - Q^{(2)}(u)\bigr)^2 \, du > 0. \tag{6}$$

Due to applications we are interested in early changes, so we assume that

$$k^* = O(1) \quad \text{if } M \to \infty. \tag{7}$$

We recall that $\sigma_N^2 = \operatorname{var}(\xi_{i,N})$, $1 \le i \le M$. If

$$\frac{N M^{1/2}}{\sigma_{M,N}} \to \infty, \tag{8}$$

then

$$\lim_{M \to \infty} P\{\tau_M < \infty\} = 1. \tag{9}$$

We note that under the conditions of Theorems 2.1, 2.2 and Remark 2.1 σ_N is bounded, so in this case (8) is satisfied. However, in the other theorems, $\sigma_N \to \infty$, so in order to have (8) we need to assume that N is a function of M and $N \to \infty$ at a certain rate. It is not difficult to establish (9) under the stated assumptions, an outline of the argument is given in Section A of the Supplementary Material. It is however not easy to derive the the asymptotic distribution of τ_M, and this may be the subject of another paper. We illustrate the distribution of τ_M in finite samples in Section 4.

3. Application to cross-sectional returns. Before we analyze finite sample properties of our procedure in Section 4, we illustrate in this section how it works in practice. We use for this purpose perhaps the most extensively studied sequence of distributions, the cross-sectional returns. Suppose $p_{i,j}$ is the price of the stock of company j at the close of trading day i. The return on day i is defined by

$$r_{i,j} = 100 \times (\log p_{i,j} - \log p_{i-1,j}).$$

The observations $r_{i,j}$, for all available companies j, are called cross-sectional returns on day i. Estimated density functions for every day in 2019 are shown in Figure 1. In finance research, cross-sectional returns over other periods, weeks, months, quarters and years have also been studied. The main strain of finance research, going back over five decades, has been concerned with determining factors which may help predict the position of the return on a stock of company j in the distribution of all returns. For example, will the value of the price to earnings ratio allow an investor to predict if the next period return will be above

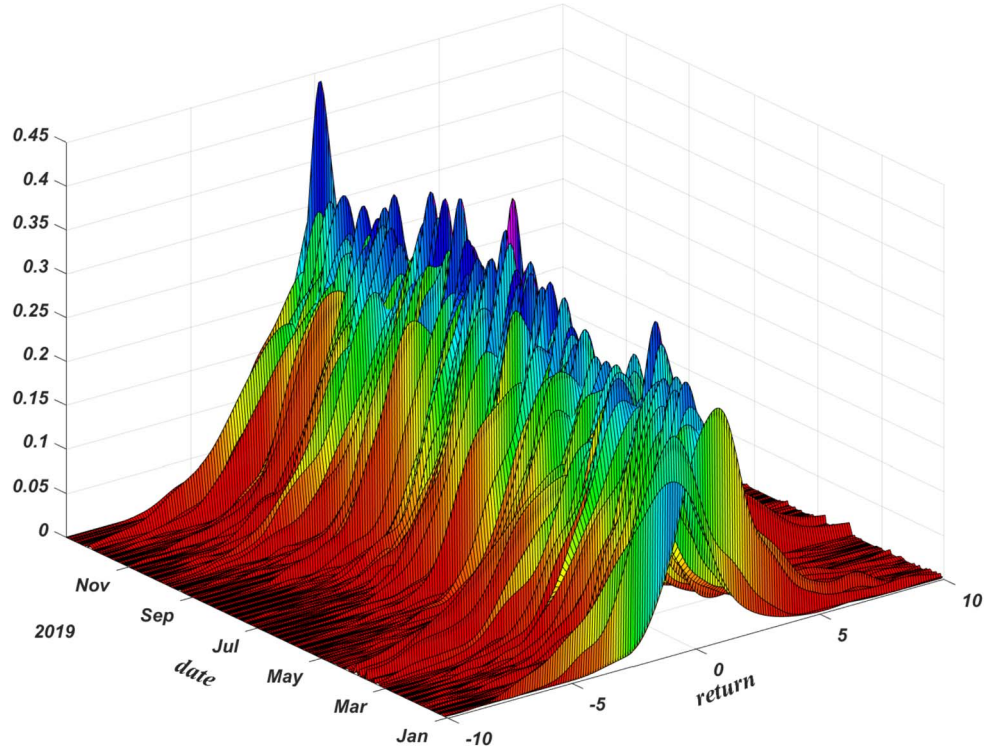

FIG. 1. *Estimated density functions of cross-sectional daily returns for stocks included in the S&P* 500 *index in* 2019. *The range of returns is in percent. It is truncated at* $\pm 10\%$ *because larger daily returns are rare and would make the graph less informative.*

the 80th quantile of all returns? Some of the most frequently cited contributions to this field include Sharpe (1964), Fama and French (1993) (26K citations as of April 2020), Carhart (1997), Ang et al. (2006) and Fama and French (2015). Consequently, the distribution of returns has been extensively studied, with a number of models proposed to describe it. Fama (1965) suggested a Gaussian mixture model, but it did not fit the data well. In the 1960s, B. Mandelbrot and E. Fama argued in favor of the stable Paretian distribution as a suitable model; see Mandelbrot (1997). Praetz (1972) and Blattberg and Gonedes (1974) advocated Student's t distribution with low degrees of freedom. Mittnik and Rachev (1993) found that the Weibull distribution gave the best fit for S&P 500 daily returns between 1982 and 1986. Granger and Ding (1995) found that the double exponential distribution is also an appropriate choice. Cont (2001) concluded that at least four parameters are needed to control the location, the scale, the skewness and the kurtosis. Chen (2005) and Wang (2012) fitted four parameter skewed t distribution to the daily cross-sectional returns of the 1000 largest capitalization stocks in the CRSP database.

The richness of models proposed for the distribution of the cross-sectional returns might be partly attributable to the expectation that this distribution might be evolving or even rapidly changing over time, and no single model can capture it over a sufficiently long period of time. This is dramatically illustrated in Figure 2, which shows p-values of the Kolmogorov–Smirnov test for the fit of a skewed t distribution, defined in Section B of the Supplementary Material, to daily cross-sectional returns of constituent stock in the S&P 500 index. We do not test if a distribution with specific parameters is a good fit, but if the whole family of distributions is suitable; these are the p-values of a goodness-of-fit test. The graph shows that the four parameter t distribution might be suitable after 1996, but would generally be a poor fit before 1996.

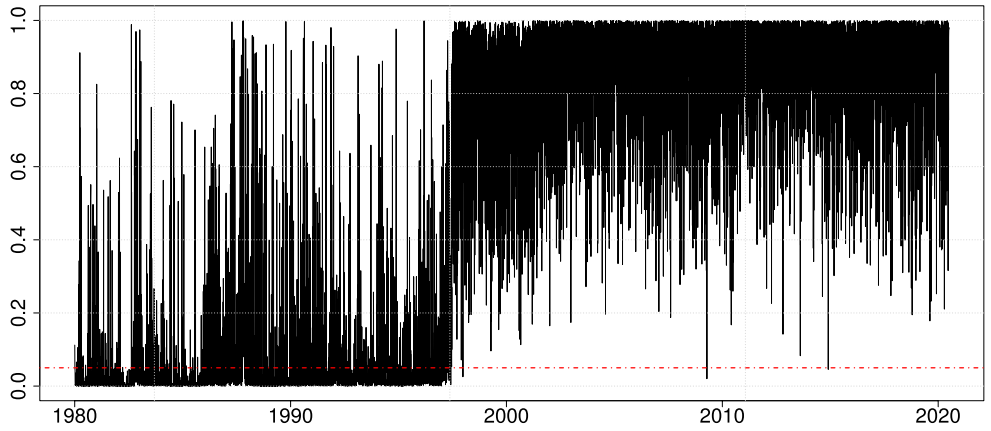

FIG. 2. *p-values of the Kolmogorov–Smirnov test for the fit of a skewed t distribution to cross-sectional returns of constituent stock in the S&P* 500 *index for every trading day from January* 1980 *until June* 2020. *The dashed line shows the* 5% *significance level.*

In this section, we focus on the application of our monitoring procedure studied in Section 2. We emphasize that it does not use any model for the distribution of cross-sectional returns and is applicable to practically any family of distributions. Our objective is to check if it can detect well-known events that impacted capital markets. We consider daily returns of the constituent stocks of the S&P 500 index, which is a widely used stock index based on the market capitalizations of about 500 largest U.S. companies. The index constituents are updated periodically according to the rules of S&P Dow Jones Indices, typically in response to acquisitions and change of market capitalizations. For the data from January 1980 to December 2019, we downloaded the historical constituents list from the UNIX server of CRSP and closing price data via the CRSP web queries. For the data between January 2020 and June 2020, we deduced the historical constituents list based on the announcements in the press release of S&P Dow Jones Indices (https://www.spglobal.com/spdji) and downloaded closing price data from Compustat. The whole dataset covers the period from January 1980 to June 2020, including 10,208 trading days.

We use the significance level $\alpha = 0.05$, $\gamma = 0.35$ and the weight function $w(t) = t(1-t)$. (The critical value from Table 1 is $c(\gamma, \alpha) = 2.4946$.) We consider four subperiods, which we identify by established names of events of impact. The detector $\Gamma(M, s)$ and the boundary function $g(M, s)$ for relevant parts of the monitoring periods are shown in Figure 3.

Subperiod 1: Black Monday in 1987 In the first subperiod, we go back to the 1980s and choose the training period of 1982–1986 (1263 trading days) and the monitoring period of 1987–1989 (758 trading days). We are interested in whether our procedure can detect the sudden and severe crash happened on Black Monday (October 19, 1987). The detector $\Gamma(M, s)$ crosses the boundary function $g(M, s)$ on October 16, 1987, three calendar days and one trading day before the Black Monday. This indicates that some realignment of stock returns started to happen before the actual crash. This can generally be determined only with a hindsight. Our sequential procedure uses only data available up to the current day.

Subperiod 2: Dot-com Bubble In the second subperiod, we shift our interest to the dotcom bubble crash and use 1995–1999 (1263 trading days) as the training period and 2000–2002 (751 trading days) as the monitoring period. The detector $\Gamma(M, s)$ crosses the boundary function on March 7, 2000. This is the date when the market peaked due to the previous years of massive growth in the use and adoption of the internet, but the market started to crash afterwards, which is referred as the burst of dot-com bubble. Our procedure can find the change in real time before the market crash.

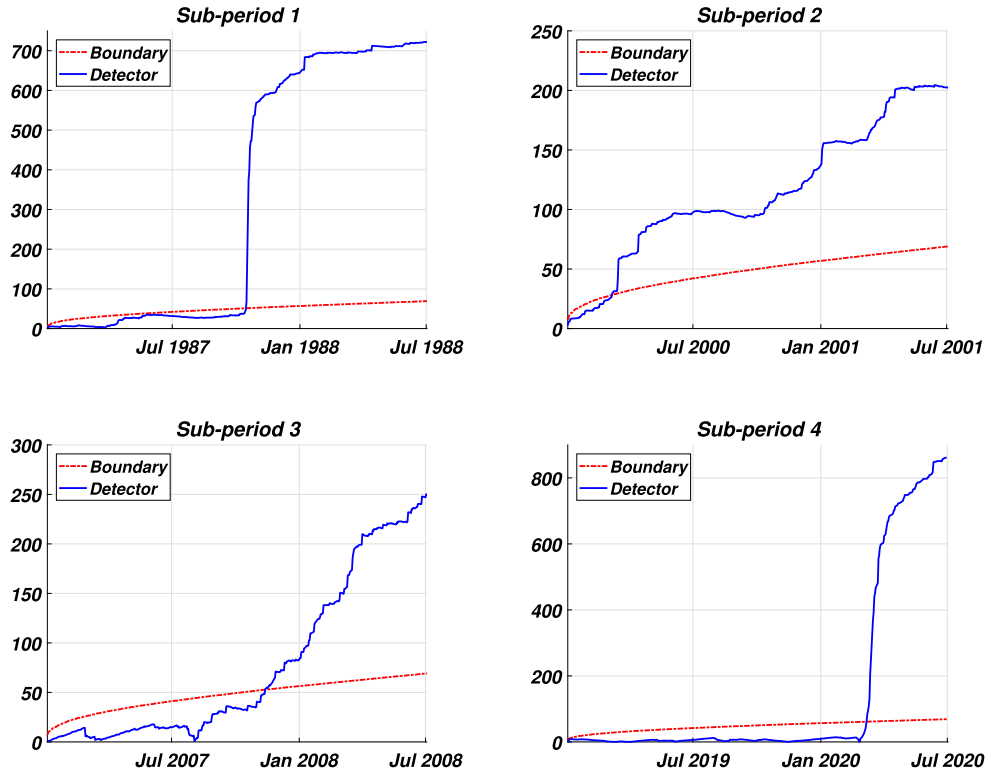

FIG. 3. *The detector $\Gamma(M, s)$ and the boundary function $g(M, s)$ for the four subperiods. Only the first one and a half years of the monitoring period is shown.*

Subperiod 3: Subprime Mortgage Crisis In the third subperiod, we focus on the global financial crisis known as the subprime mortgage crisis of 2008. The training period and the monitoring period are set to be 2003–2006 (1007 trading days) and 2007–2010 (756 trading days), respectively. The Wasserstein distances $\xi_{i,N}$ (not shown) initially stays at a low level, but it starts to climb up in the late 2007 and has substantial growth since the financial crisis in 2008. The increase of $\xi_{i,N}$ in the monitoring period is precisely reflected in the detector $\Gamma(M, s)$, which exceeds the boundary function on November 13, 2007. This is several months before September 2008 when the Subprime Mortgage Crisis began to be apparent (on September 15, 2008, Lehman Brothers filed for bankruptcy). This example indicates that our procedure can provide early warnings.

Subperiod 4: COVID-19 In the last subperiod, we are interested in monitoring the dramatic impact from COVID-19. We choose the test period from January 2019 to June 2020 (377 trading days) and use the previous 5 years (2014–2018, 1257 trading days) as the training period. The detector $\Gamma(M, s)$ abruptly goes over the boundary function $g(M, s)$ on March 9, 2020. Although it is in the early stage of the pandemic in the U.S., the S&P 500 index dropped 7% within 3 minutes after the market opening on that day and triggered the circuit breaker, resulting in the trading halt of all stocks. This is the first time that the circuit breaker is triggered in the last two decades, and then it is triggered for three additional times in March 2020. Our procedure can promptly detect the abrupt change.

There could be a concern that raw returns may not satisfy the independence conditions specified in Assumption 2.1. As a robustness check, we applied our procedure to idiosyncratic returns (IR). An idiosyncratic return is defined as the fraction of the excess return not explained by common factors; Morgenson and Harvey (2002) provide an introduction. For this reason, idiosyncratic returns can be considered as "more independent" than raw re-

TABLE 2
Time of detected change based on raw return and idiosyncratic return

	Subperiod 1	Subperiod 2	Subperiod 3	Subperiod 4
Raw Returns	Oct 16, 1987	Mar 7, 2000	Nov 13, 2007	Mar 9, 2020
IR from 1-factor CAPM	Jun 16, 1987	Jan 3, 2000	Aug 10, 2007	Mar 3, 2020
IR from 3-factor Model	Jun 15, 1987	Jan 3, 2000	Aug 10, 2007	Mar 10, 2020
IR from 4-factor Model	Jun 2, 1987	Jan 3, 2000	Nov 12, 2007	Mar 9, 2020
IR from 5-factor Model	Jun 2, 1987	Jan 3, 2000	Aug 13, 2007	Mar 10, 2020

turns. To construct idiosyncratic returns, a factor model for stock returns is needed to account for various common risk factors. In order to explore the potential impact from the specific choice of a factor model, we decide to use four different factor models, including the one-factor capital asset pricing model (CAPM), the three-factor model (Fama and French (1993)), the four-factor model (Carhart (1997)) and the five-factor model (Fama and French (2015)). Then we follow the specific method of Herskovic et al. (2016) to obtain idiosyncratic returns in the training period. Since our procedure monitors the change in real-time, the idiosyncratic returns in the monitoring period are constructed based on factor loadings from previous year. The factors were downloaded from Professor Kenneth R. French's website: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Using the same setting, we perform our monitoring procedure on the idiosyncratic returns of constituent stocks in the S&P 500 index. Table 2 compares the time of detected changes based on the raw returns and the idiosyncratic returns. As can be observed, the results based on the idiosyncratic returns from four different factor models are generally similar, implying that the choice of the specific factor model has small impact. In the first three subperiods, we can typically detect changes a few months earlier based on the idiosyncratic returns, compared with the raw return. In Sub-period 4, there is no substantial difference in the time of detected change, based on the raw and idiosyncratic returns. It appears that our procedure, generally, has shorter delay time if the independence assumption is more closely satisfied by the data.

In the above examples, the length of the training period, M, was chosen in such a way that it does not contain any obvious events that might violate the assumption of a constant distribution of cross-sectional returns. This period must also be sufficiently long because the validity of our methods is established as $M \to \infty$. A data driven method of choosing the longest past period of stable returns might be developed, as was done Chen, Härdle and Pigorsch (2010) in the case of a scalar series of realized volatility. It is however not clear at present how to do it for quantile functions, and a reasonable exploratory analysis might provide a superior choice. The length of the monitoring period, K, was chosen for illustration only. Since our theory assumes that $M + K \to \infty$, K can be chosen to be fairly long. In fact, except for subperiod 4, for which we had no more data, we used K which is several times longer than the one and a half year period used in Figure 3. Since the boundary crossing occurred within one and a half year after monitoring commenced, we displayed only the first one and a half year of monitoring so that the boundary function and the detector can be seen more clearly. The training period would be updated periodically, definitely after a change has been detected. Once a change is detected, the procedure is terminated and can be restarted after a sufficiently long training sample is available.

4. A simulation study. For clarity of presentations, we start with the definitions of the quantities appearing in this section:

- N is the number of observations at each time i.
- M is the length of historical training sample.
- K is the length of monitoring sample.
- $w(t)$ is the weight function in (1).
- γ is the exponent in (3).
- α is the size in (4).

In the examples of Section 3, we used $N \approx 500$, $M \approx 1250$, $K \approx 750$, $w(t) = t(1-t)$, $\gamma = 0.35$, $\alpha = 0.05$. In this section, we explore the behavior of the procedure for broader ranges of these values. Before doing so, we summarize the procedure in an algorithmic form, which may be useful for researchers who want to apply it without studying the underlying theory.

Steps of the detection procedure:

1. At each time i, order the data from smallest to largest, $X_{i;1,N} \le X_{i;2,N} \le \cdots \le X_{i;N,N}$.
2. Calculate the quantile function

$$Q_{i,N}(t) = X_{i;j,N} \quad \text{if } (j-1)/(N+1) \le t < j/(N+1),\ j = 1, 2, \ldots, N.$$

The values $Q_{i,N}(t)$ are numerically calculated at the grid $t \in \{1/2N, 2/2N, \ldots, 1 - 1/2N\}$.

3. Calculate the average quantile function of the training sample:

$$\bar{Q}_M(t) = \frac{1}{M}\sum_{i=1}^{M} Q_{i,N}(t).$$

4. Calculate the distances

$$\xi_{i,N} = \int_{1/(2N)}^{1-1/(2N)} \{Q_{i,N}(t) - \bar{Q}_M(t)\}^2 w(t)\,dt, \quad 1 \le i \le M+K.$$

The integral is calculated numerically over the grid $t \in \{1/2N, 2/2N, \ldots, 1 - 1/2N\}$.

5. Calculate the mean and the variance of the $\xi_{i,N}$ in the training sample:

$$\hat{\xi}_{M,N} = \frac{1}{M}\sum_{i=1}^{M} \xi_{i,N} \quad \text{and} \quad \hat{\sigma}^2_{M,N} = \frac{1}{M-1}\sum_{i=1}^{M} (\xi_{i,N} - \hat{\xi}_{M,N})^2.$$

6. Calculate the detector $\Gamma(M, s)$,

$$\Gamma(M,s) = \frac{1}{\hat{\sigma}_{M,N}} \left| \sum_{i=M+1}^{M+s} \xi_{i,N} - \frac{s}{M}\sum_{i=1}^{M} \xi_{i,N} \right|, \quad 1 \le s \le K.$$

7. Calculate boundary function

$$g(M,s) = c(\gamma,\alpha) M^{1/2} \left(1 + \frac{s}{M}\right)\left(\frac{s}{s+M}\right)^{\gamma}, \quad 1 \le s \le K,$$

where $c(\gamma, \alpha)$ is from Table 1.

8. Reject H_0 if there is a $1 \le s \le K$ such that $\Gamma(M, s) > g(M, s)$. The first s at which this occurs is the time of detection.

The above steps are repeated 5000 times to compute empirical size and power for a specific data generating process $X_{i,j}$, $1 \le j \le N$, $i = 1, 2, \ldots, M+K$. There are basically unlimited choices for the distribution of the $X_{i,j}$, the type of change, and for the values of N, M and K. We consider several scenarios, using the data analysis in Section 3 as a motivation.

Empirical size We consider two data generating processes (DGPs): (1) skewed t distribution with $\mu = 0$, $\sigma = 2$, $\xi = 1.05$, $\nu = 4.5$; (2) standard normal distribution. The skewed

t distribution is motivated by the application to the cross-sectional returns in Section 3. Its density and parameters are specified in Section B of the Supplementary Material. The parameter values we use are representative to what we have seen after year 2000 in relatively stable periods. The normal distribution is used to provide a more generic example, which may be relevant in many other fields. We consider two lengths of the historical training sample, $M = 500$ and 1250, and three lengths of the monitoring period $K = 750$, 1250 and 2500. There are $N = 500$ observations at each time. The weight function is chosen to be $w(t) = t(1-t)$. The asymptotic theory justifies the procedure for $K = \infty$, so we expect the sizes to be generally smaller than nominal for finite K (the crossing may occur later).

Table 3 reports the empirical sizes for the monitoring scheme for two DGPs at the significance levels of 1%, 5% and 10% with a range of values of γ. The monitoring procedure has reasonably good empirical sizes because they are generally under (when K is small or modest) or close to (when K is large) the nominal sizes as suggested by the theory developed in Section 2. Additionally, the empirical sizes for $M = 1250$ are closer to theoretical levels, which also reflects the asymptotic validity as $M \to \infty$. One noticeable insight is that the empirical sizes depend on the choice of γ. Recall that γ can be arbitrarily chosen between 0 and 0.5. The observations in our simulation shows that a larger γ results in a higher rejection percentage and a smaller γ is more conservative in rejection. Specifically, the nominal sizes are close to theoretical levels but with marginal inflation if γ is close to 0.5, while opposite direction is observed if γ is close to 0. Considering the case most closely related to the empirical study in Section 3 (skewed t, $\gamma = 0.35$, $M = 1250$, $k = 750$), we see that the probability of a type I error (false detection) is about 2.1%. Thus the detections reported in the first three subperiods considered in Section 3 are unlikely to be spurious. (They reflect well-known real events, which is a stronger justification.) We have also computed the rejection rates for the weight function $w(t) = 1$, which corresponds to the usual Kantorovich–Wasserstein distance. It gives good sizes for the normal distribution, but too many rejections for the skewed t distribution; a constant weight function apparently places too much weight on the heavy tails.

Finally, we also considered the case of temporally dependent observations. Even though this case is not covered by our theory, it is useful to see if the method still performs well if the assumption of independence is violated. We generated data according to GARCH(1, 1) model:

$$X_{i,j} = \sigma_{i,j}\varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim i.i.d.\mathcal{N}(0,1),$$
$$\sigma_{i,j}^2 = 0.05 + 0.9\sigma_{i-1,j}^2 + 0.05\varepsilon_{i-1,j}^2.$$

The parameter values are very typical of what is encountered for real stocks. The empirical size is only marginally higher than for independent data, and the power is correspondingly marginally higher.

Empirical power We now turn to the analysis of the empirical power. There are many different ways of changes under the alternative, especially for the skewed t distribution. In reality, we typically observe a larger volatility, a more negative skewness, and a higher kurtosis in the distribution of cross-sectional returns in presence of a crisis. The opposite direction occurs when the economy recovers from a crisis. Thus, we consider both positive and negative changes in different moments of the skewed t distribution in our simulation study under the alternative. In addition, we explore two scenarios related to the time of change k^*: (*i*) $k^* = 1$ represents the scenario when the distribution of the observations changes immediately after the training sample; (*ii*) $k^* = 100$ implies that the changes occurs relatively late in the monitoring sample. The historical sample length is $M = 750$ and 1250, the monitoring sample is $K = 750$, and there are $N = 500$ scalar observations at each time. The weight

TABLE 3
Empirical size with weight function $w(t) = t(1-t)$

	DGP-1: skewed t ($\mu = 0, \sigma = 2, \xi = 1.05, \nu = 4.5$)								
	$M = 500$, $K = 750$			$M = 500$, $K = 1250$			$M = 500$, $K = 2500$		
$\gamma \setminus \alpha$	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.00	0.3%	1.5%	3.1%	0.7%	2.4%	5.1%	1.1%	4.0%	7.5%
0.15	0.5%	2.2%	4.7%	0.9%	3.1%	6.6%	1.4%	4.4%	8.6%
0.25	0.6%	2.9%	6.0%	1.1%	3.7%	7.6%	1.5%	4.9%	9.2%
0.35	1.3%	4.4%	8.0%	1.6%	5.0%	9.2%	1.9%	5.8%	10.1%
0.45	2.9%	7.0%	10.6%	3.2%	7.4%	11.1%	3.4%	7.7%	11.6%
0.49	3.2%	7.0%	10.2%	3.3%	7.3%	10.5%	3.4%	7.5%	10.7%
	$M = 1250$, $K = 750$			$M = 1250$, $K = 1250$			$M = 1250$, $K = 2500$		
$\gamma \setminus \alpha$	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.00	0.0%	0.1%	0.4%	0.0%	0.5%	1.5%	0.1%	1.7%	3.8%
0.15	0.0%	0.4%	1.2%	0.1%	1.1%	2.5%	0.2%	2.4%	4.9%
0.25	0.2%	0.9%	2.2%	0.3%	1.8%	3.6%	0.4%	3.1%	5.9%
0.35	0.3%	2.1%	4.4%	0.4%	2.9%	5.6%	0.7%	4.0%	7.4%
0.45	2.1%	5.3%	9.1%	2.2%	5.9%	10.1%	2.3%	6.5%	11.1%
0.49	2.9%	6.4%	9.7%	2.9%	6.7%	10.2%	2.9%	6.9%	10.8%
	DGP-2: standard normal								
	$M = 500$, $K = 750$			$M = 500$, $K = 1250$			$M = 500$, $K = 2500$		
$\gamma \setminus \alpha$	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.00	0.4%	1.5%	3.0%	0.7%	2.7%	5.4%	1.1%	4.1%	7.3%
0.15	0.6%	2.3%	4.3%	0.9%	3.6%	6.6%	1.4%	4.8%	8.0%
0.25	0.9%	3.2%	5.8%	1.1%	4.3%	7.8%	1.6%	5.4%	8.9%
0.35	1.5%	4.6%	7.9%	1.7%	5.6%	9.5%	2.1%	6.3%	10.5%
0.45	3.3%	7.3%	10.7%	3.5%	7.8%	11.6%	3.7%	8.1%	12.2%
0.49	3.7%	7.5%	10.8%	3.8%	7.8%	11.2%	3.9%	8.0%	11.5%
	$M = 1250$, $K = 750$			$M = 1250$, $K = 1250$			$M = 1250$, $K = 2500$		
$\gamma \setminus \alpha$	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.00	0.0%	0.1%	0.5%	0.0%	0.5%	1.3%	0.2%	1.6%	3.6%
0.15	0.0%	0.5%	1.4%	0.1%	1.0%	2.6%	0.4%	2.3%	4.8%
0.25	0.1%	1.2%	2.8%	0.4%	1.8%	4.2%	0.6%	3.0%	6.0%
0.35	0.6%	2.5%	4.7%	0.8%	3.1%	6.0%	1.0%	4.2%	7.5%
0.45	2.1%	5.8%	8.7%	2.3%	6.2%	9.5%	2.4%	6.8%	10.3%
0.49	2.8%	6.9%	9.8%	3.0%	7.2%	10.1%	3.0%	7.4%	10.6%

function is $w(t) = t(1-t)$. We set the values of parameters of skewed t before k^* as

$$X_{i,j} \sim \text{ skewed } t\ (\mu = \mu_0, \sigma = \sigma_0, \xi = \xi_0, \nu = \nu_0), \quad 1 \le i < M + k^*, 1 \le j \le N,$$

where $\mu_0 = 0$, $\sigma_0 = 2$, $\xi_0 = 1.05$ or 0.95, $\nu_0 = 0.45$ or 0.55, and we change the values in one of four parameters in each of the following alternatives:

$$H_{A,1} : X_{i,j} \sim \text{ skewed } t\ (\mu = \mu_1, \sigma = \sigma_0, \xi = \xi_0, \nu = \nu_0), \quad M + k^* \le i \le M + K,$$

$$H_{A,2} : X_{i,j} \sim \text{ skewed } t\ (\mu = \mu_0, \sigma = \sigma_1, \xi = \xi_0, \nu = \nu_0), \quad M + k^* \le i \le M + K,$$

$$H_{A,3} : X_{i,j} \sim \text{ skewed } t\ (\mu = \mu_0, \sigma = \sigma_0, \xi = \xi_1, \nu = \nu_0), \quad M + k^* \le i \le M + K,$$

$$H_{A,4} : X_{i,j} \sim \text{ skewed } t\ (\mu = \mu_0, \sigma = \sigma_0, \xi = \xi_0, \nu = \nu_1), \quad M + k^* \le i \le M + K.$$

TABLE 4

Empirical power under $H_{A,1}$ and $H_{A,2}$ with weight function $w(t) = t(1-t)$ at significance level 5%

	$H_{A,1}$: skewed t ($\mu = \mu_1, \sigma = 2, \xi = 1.05, \nu = 4.5$)				$H_{A,2}$: skewed t ($\mu = 0, \sigma = \sigma_1, \xi = 1.05, \nu = 4.5$)			
Before k^*:	$\mu_0 = 0.00$	$\mu_0 = 0.00$	$\mu_0 = 0.00$	$\mu_0 = 0.00$	$\sigma_0 = 2.00$	$\sigma_0 = 2.00$	$\sigma_0 = 2.00$	$\sigma_0 = 2.00$
After k^*:	$\mu_1 = -0.10$	$\mu_1 = -0.05$	$\mu_1 = 0.05$	$\mu_1 = 0.10$	$\sigma_1 = 1.80$	$\sigma_1 = 1.90$	$\sigma_1 = 2.10$	$\sigma_1 = 2.20$
γ	$M = 500, k^* = 1$				$M = 500, k^* = 1$			
0.00	100.0%	91.4%	90.8%	100.0%	100.0%	79.4%	100.0%	100.0%
0.15	100.0%	93.4%	92.8%	100.0%	100.0%	83.1%	100.0%	100.0%
0.25	100.0%	94.2%	93.8%	100.0%	100.0%	85.0%	100.0%	100.0%
0.35	100.0%	94.7%	94.4%	100.0%	100.0%	85.7%	100.0%	100.0%
0.45	100.0%	94.3%	93.7%	100.0%	100.0%	83.7%	100.0%	100.0%
0.49	100.0%	92.3%	91.8%	100.0%	100.0%	79.3%	100.0%	100.0%
γ	$M = 1250, k^* = 1$				$M = 1250, k^* = 1$			
0.00	100.0%	94.6%	94.2%	100.0%	100.0%	82.1%	99.9%	100.0%
0.15	100.0%	97.4%	97.2%	100.0%	100.0%	90.2%	99.9%	100.0%
0.25	100.0%	98.3%	98.4%	100.0%	100.0%	93.2%	99.9%	100.0%
0.35	100.0%	98.6%	98.7%	100.0%	100.0%	95.0%	99.9%	100.0%
0.45	100.0%	98.8%	98.8%	100.0%	100.0%	95.2%	99.9%	100.0%
0.49	100.0%	98.4%	98.5%	100.0%	100.0%	93.5%	99.9%	100.0%
γ	$M = 500, k^* = 100$				$M = 500, k^* = 100$			
0.00	100.0%	81.9%	81.8%	100.0%	100.0%	65.5%	100.0%	100.0%
0.15	100.0%	84.5%	84.6%	100.0%	100.0%	69.7%	100.0%	100.0%
0.25	100.0%	85.7%	85.9%	100.0%	100.0%	72.1%	100.0%	100.0%
0.35	100.0%	86.1%	86.1%	100.0%	100.0%	72.4%	100.0%	100.0%
0.45	100.0%	83.6%	83.7%	100.0%	100.0%	68.3%	100.0%	100.0%
0.49	100.0%	79.6%	79.1%	100.0%	100.0%	61.4%	100.0%	100.0%
γ	$M = 1250, k^* = 100$				$M = 1250, k^* = 100$			
0.00	100.0%	86.1%	84.2%	100.0%	100.0%	64.5%	100.0%	100.0%
0.15	100.0%	92.2%	91.0%	100.0%	100.0%	77.2%	100.0%	100.0%
0.25	100.0%	94.7%	93.6%	100.0%	100.0%	83.1%	100.0%	100.0%
0.35	100.0%	95.9%	95.0%	100.0%	100.0%	86.2%	100.0%	100.0%
0.45	100.0%	96.0%	94.9%	100.0%	100.0%	86.2%	100.0%	100.0%
0.49	100.0%	94.4%	93.7%	100.0%	100.0%	82.5%	100.0%	100.0%

Tables 4 and 5 show the empirical power for the above four alternatives for $k^* = 1$ and 100 at the significance level of 5%. There are five major observations. First, our test has higher power when the change occurs relatively early in the monitoring sample since the rejection rates of $k^* = 1$ are generally higher than those of $k^* = 100$. Second, since the power tends to one as $M \to \infty$, the empirical power of $M = 1250$ is, as predicted by the theory, higher than $M = 500$ in most cases. Third, the power approaches 100% as the magnitude of the change increases in either positive or negative directions. Fourth, the choice of γ has a minor impact on the power; it usually peaks if $\gamma = 0.35$ or 0.45. Lastly, the power has symmetric pattern only in $H_{A,1}$, while the test is more sensitives to change of increase in σ, ξ, and ν.

In Section C of the Supplementary Material, we consider the impact of the weight function w on the performance of our method. A broad conclusion is that while for distributions with light tails the choice of w does not matter much, for heavy-tailed observations functions giving less weights to tails are recommended. This emphasizes the usefulness of using weighted quantile functions rather than the original Wasserstein distance.

TABLE 5

Empirical power under $H_{A,3}$ and $H_{A,4}$ with weight function $w(t) = t(1-t)$ at significance level 5%

	$H_{A,3}$: skewed t ($\mu = 0, \sigma_0 = 2, \xi = \xi_1, \nu = 4.5$)				$H_{A,4}$: skewed t ($\mu = 0, \sigma_0 = 2, \xi = 1.05, \nu = \nu_1$)			
Before k^*:	$\xi_0 = 1.05$	$\xi_0 = 1.05$	$\xi_0 = 0.95$	$\xi_0 = 0.95$	$\nu_0 = 4.50$	$\nu_0 = 4.50$	$\nu_0 = 5.50$	$\nu_0 = 5.50$
After k^*:	$\xi_1 = 0.95$	$\xi_1 = 0.99$	$\xi_1 = 1.01$	$\xi_1 = 1.05$	$\nu_1 = 5.80$	$\nu_1 = 5.50$	$\nu_1 = 4.50$	$\nu_1 = 4.20$
γ	$M = 500, k^* = 1$				$M = 500, k^* = 1$			
0.00	100.0%	73.9%	80.5%	100.0%	100.0%	96.9%	77.6%	100.0%
0.15	100.0%	78.3%	84.3%	100.0%	100.0%	97.7%	81.2%	100.0%
0.25	100.0%	80.4%	86.0%	100.0%	100.0%	98.1%	83.4%	100.0%
0.35	100.0%	81.2%	86.7%	100.0%	100.0%	98.3%	84.2%	100.0%
0.45	100.0%	79.6%	85.5%	100.0%	100.0%	97.9%	82.0%	100.0%
0.49	100.0%	74.6%	80.9%	100.0%	99.9%	96.7%	77.2%	100.0%
γ	$M = 1250, k^* = 1$				$M = 1250, k^* = 1$			
0.00	100.0%	74.6%	82.9%	100.0%	100.0%	98.8%	79.4%	100.0%
0.15	100.0%	85.0%	91.1%	100.0%	100.0%	99.5%	89.0%	100.0%
0.25	100.0%	89.3%	94.2%	100.0%	100.0%	99.7%	92.7%	100.0%
0.35	100.0%	92.0%	95.7%	100.0%	100.0%	99.8%	94.3%	100.0%
0.45	100.0%	92.4%	96.0%	100.0%	100.0%	99.9%	94.7%	100.0%
0.49	100.0%	90.3%	94.6%	100.0%	100.0%	99.8%	93.1%	100.0%
γ	$M = 500, k^* = 100$				$M = 500, k^* = 100$			
0.00	100.0%	61.2%	66.9%	100.0%	99.6%	90.9%	64.0%	99.9%
0.15	100.0%	65.4%	70.9%	100.0%	99.7%	92.7%	68.4%	99.9%
0.25	100.0%	67.4%	72.9%	100.0%	99.7%	93.6%	70.3%	100.0%
0.35	100.0%	67.9%	73.6%	100.0%	99.7%	93.6%	70.8%	100.0%
0.45	100.0%	64.7%	70.1%	100.0%	99.7%	91.9%	67.0%	99.9%
0.49	100.0%	57.6%	64.1%	100.0%	99.5%	89.1%	60.0%	99.9%
γ	$M = 1250, k^* = 100$				$M = 1250, k^* = 100$			
0.00	100.0%	58.3%	66.2%	100.0%	100.0%	94.4%	62.0%	100.0%
0.15	100.0%	71.5%	78.1%	100.0%	100.0%	97.5%	75.1%	100.0%
0.25	100.0%	78.3%	83.3%	100.0%	100.0%	98.4%	81.8%	100.0%
0.35	100.0%	81.9%	86.9%	100.0%	100.0%	98.8%	84.9%	100.0%
0.45	100.0%	81.9%	86.8%	100.0%	100.0%	98.7%	85.0%	100.0%
0.49	100.0%	78.2%	83.1%	100.0%	100.0%	98.2%	81.4%	100.0%

It is also worthwhile to examine the distribution of the stopping time τ_M. Figure 4 shows the estimated densities of the distributions of the τ_M under scenarios focusing two different time of change $k^* = 1$ (upper panel) and $k^* = 100$ (lower panel), under $H_{A,2}$ with $\sigma_1 = 2.1$, $M = 500$ and $K = 750$. The detectors with γ close to 0.5 have the shortest delay in detection in the scenario which a change occurs immediately in the monitoring sample, while the detector with $\gamma = 0$ typically find the first exceedance above the boundary function with delay in a period of 118 (mode in its density). However, the detector with $\gamma = 0.45$ or 0.49 is not recommended for the practical application due to the insight found in the lower panel of Figure 4 for the scenario of $k^* = 100$, where we observe some spurious detections even before a change has occurred. Overall, $\gamma = 0.35$ (the solid thick line in Figure 4) is our recommended choice because it gives a good balance between the short delay and the negligible proportion of false early alarms. Section D of the Supplementary Material provides additional figures illustrating the distribution of the stopping time in different scenarios; $\gamma = 0.35$ continuous

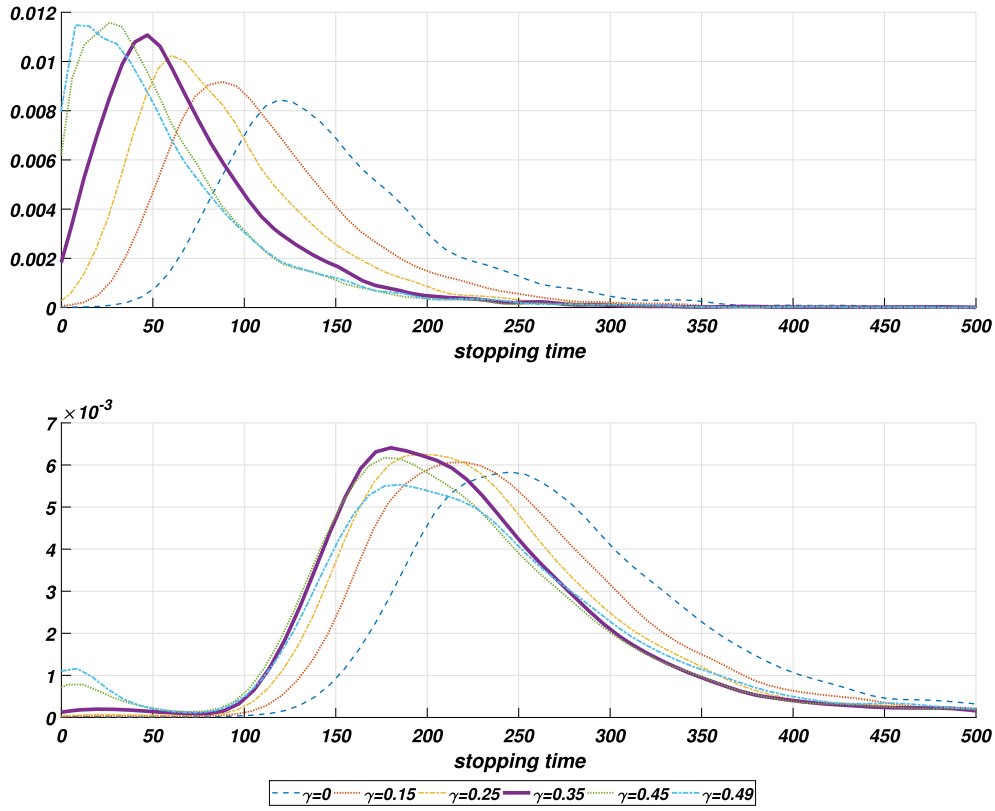

FIG. 4. *Estimated densities of the stopping time of rejection of* $H_{A,2}$ *at the* 5% *significance level. Upper panel*: $k^* = 1$. *Lower Panel*: $k^* = 100$.

to be the preferred choice. We also briefly discuss in Section D what happens if there are two change points close to each other.

Our theory covers both the case of fixed N and $N \to \infty$. A practical question is how small can N be? Recall that N is the number of observations needed to estimate the quantile function (or cdf or density), so this number cannot be too small. Additional simulations, not reported, show that using $N = 100$ rather than $N = 500$ has no noticeable impact on the empirical size and reduces the power slightly.

Acknowledgments. We thank two referees for posing substantive questions and providing useful advice, which have lead to a much improved paper.

Funding. This research was partially supported by NSF Grant DMS–1914882.

SUPPLEMENTARY MATERIAL

Supplementary material for "Monitoring for a change point in a sequence of distributions" (DOI: 10.1214/20-AOS2036SUPP; .pdf). In Section A, we provide the proofs of the results of Section 2. In Section B, we show the parameterization of skewed t distribution. In Section C, we investigate the impact of the weight function on empirical rejection rates. In Section D, we illustrate the distribution of the stopping time in different scenarios and briefly discuss what happens if there are two change points close to each other.

REFERENCES

ANG, A., HODRICK, R. J., XING, Y. and ZHANG, X. (2006). The cross-section of volatility and expected returns. *J. Finance* **61** 259–299.

BARDSLEY, P., HORVÁTH, L., KOKOSZKA, P. and YOUNG, G. (2017). Change point tests in functional factor models with application to yield curves. *Econom. J.* **20** 86–117. MR3636962 https://doi.org/10.1111/ectj.12075

BARIGOZZI, M., CHO, H. and FRYZLEWICZ, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *J. Econometrics* **206** 187–225. MR3840788 https://doi.org/10.1016/j.jeconom.2018.05.003

BERTHET, P. and FORT, J. C. (2020). Exact rate of convergence of the expected W_2 distance between the empirical and true Gaussian distribution. *Electron. J. Probab.* **25** Paper No. 12, 16. MR4059190 https://doi.org/10.1214/19-ejp410

BINGHAM, N. H., GOLDIE, C. M. and TEUGELS, J. L. (1987). *Regular Variation. Encyclopedia of Mathematics and Its Applications* **27**. Cambridge Univ. Press, Cambridge. MR0898871 https://doi.org/10.1017/CBO9780511721434

BLATTBERG, R. C. and GONEDES, N. J. (1974). A comparison of the stable and student distributions as statistical models for stock prices. *J. Bus.* **47** 244–280.

CARHART, M. M. (1997). On persistence in mutual fund performance. *J. Finance* **52** 57–82.

CHEN, C. (2005). *Four Essays on the Equity Market*: *Decimalization*, *Volatility*, *Indexation and Cross-Sectional Returns*. ProQuest LLC, Ann Arbor, MI. Ph.D. thesis, Univ. Illinois, Chicago, IL. MR2708219

CHEN, Y., HÄRDLE, W. K. and PIGORSCH, U. (2010). Localized realized volatility modeling. *J. Amer. Statist. Assoc.* **105** 1376–1393. MR2796557 https://doi.org/10.1198/jasa.2010.ap09039

CHEN, Y., WANG, T. and SAMWORTH, R. J. (2020). High-dimensional, multiscale online changepoint detection. Technical Report, Univ. Cambridge. arXiv:2003.03668.

CHU, C.-S. J., STINCHCOMBE, M. and WHITE, H. (1996). Monitoring structural change. *Econometrica* **64** 1045–1065.

CONT, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223–236.

CSÖRGŐ, M. and HORVÁTH, L. (1993). *Weighted Approximations in Probability and Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. With a foreword by David Kendall. MR1215046

CSÖRGŐ, M. and RÉVÉSZ, P. (1978). Strong approximations of the quantile process. *Ann. Statist.* **6** 882–894. MR0501290

CSÖRGŐ, M. and RÉVÉSZ, P. (1981). *Strong Approximations in Probability and Statistics*. *Probability and Mathematical Statistics*. Academic Press, New York. MR0666546

DEL BARRIO, E., GINÉ, E. and MATRÁN, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27** 1009–1071. MR1698999 https://doi.org/10.1214/aop/1022677394

DEL BARRIO, E., GINÉ, E. and UTZET, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189. MR2121458 https://doi.org/10.3150/bj/1110228245

DEL BARRIO, E., CUESTA-ALBERTOS, J. A., MATRÁN, C. and RODRÍGUEZ-RODRÍGUEZ, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.* **27** 1230–1239. MR1740113 https://doi.org/10.1214/aos/1017938923

DUBEY, P. and MÜLLER, H.-G. (2020). Fréchet change-point detection. *Ann. Statist.* **48** 3312–3335. MR4185810 https://doi.org/10.1214/19-AOS1930

FAMA, E. (1965). The behavior of stock-market prices. *J. Bus.* **38** 34–105.

FAMA, E. and FRENCH, K. (1993). Common risk factors in the returns on bonds and stocks. *J. Financ. Econ.* **33** 3–56.

FAMA, E. and FRENCH, K. (2015). A five-factor asset pricing model. *J. Financ. Econ.* **116** 1–22.

GRANGER, C. W. J. and DING, Z. (1995). Some properties of absolute return: An alternative measure of risk. *Ann. Econ. Statist.* **40** 67–91. MR1476513 https://doi.org/10.2307/20076016

GROMENKO, O., KOKOSZKA, P. and REIMHERR, M. (2017). Detection of change in the spatiotemporal mean function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 29–50. MR3597963 https://doi.org/10.1111/rssb.12156

HERSKOVIC, B., KELLY, B., LUSTIG, H. and VAN NIEUWERBURGH, S. (2016). The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *J. Financ. Econ.* **119**.

HORVÁTH, L., KOKOSZKA, P. and WANG, S. (2021). Supplement to "Monitoring for a change point in a sequence of distributions." https://doi.org/10.1214/20-AOS2036SUPP

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. MR3405600 https://doi.org/10.1214/15-AOS1347

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. MR0375412 https://doi.org/10.1007/BF00533093

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrsch. Verw. Gebiete* **34** 33–58. MR0402883 https://doi.org/10.1007/BF00532688

LI, J. and JIN, B. (2018). Multi-threshold accelerated failure time model. *Ann. Statist.* **46** 2657–2682. MR3851751 https://doi.org/10.1214/17-AOS1632

MANDELBROT, B. (1997). *The Variation of Certain Speculative Prices*. Springer, Berlin.

MITTNIK, S. and RACHEV, S. T. (1993). Modeling asset returns with alternative stable distributions. *Econometric Rev.* **12** 261–389. With comments by P. C. B. Phillips, F. X. Diebold and R. T. Baillie and a reply by the authors. MR1249625 https://doi.org/10.1080/07474939308800266

MORGENSON, G. and HARVEY, C. (2002). *The New York Times Dictionary of Money and Investing*: *The Essential A-to-Z Guide to the Language of the New Market*. Macmillan, New York.

PADILLA, O. H. M., ATHEY, A., REINHART, A. and SCOTT, J. G. (2019). Sequential nonparametric tests for a change in distribution: An application to detecting radiological anomalies. *J. Amer. Statist. Assoc.* **114** 514–528. MR3963159 https://doi.org/10.1080/01621459.2018.1476245

PANARETOS, V. M. and ZEMEL, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.* **44** 771–812. MR3476617 https://doi.org/10.1214/15-AOS1387

PANARETOS, V. M. and ZEMEL, Y. (2019). Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **6** 405–431. MR3939527 https://doi.org/10.1146/annurev-statistics-030718-104938

PETERSEN, A. and MÜLLER, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Statist.* **44** 183–218. MR3449766 https://doi.org/10.1214/15-AOS1363

POLLAK, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13** 206–227. MR0773162 https://doi.org/10.1214/aos/1176346587

POLUNCHENKO, A. S. and TARTAKOVSKY, A. G. (2010). On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *Ann. Statist.* **38** 3445–3457. MR2766858 https://doi.org/10.1214/09-AOS775

PRAETZ, P. (1972). The distribution of share price changes. *J. Bus.* **45** 49–55.

SHARPE, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance* **19** 425–442.

SKOROKHOD, A. V. (1965). *Studies in the Theory of Random Processes*. Addison-Wesley, Reading, MA. Translated from the Russian by Scripta Technica, Inc. MR0185620

WANG, J. (2012). *A State Space Model Approach to Functional Time Series and Time Series Driven by Differential Equations*. ProQuest LLC, Ann Arbor, MI. Ph.D. thesis, Rutgers Univ., New Brunswick, NJ. MR3152378

XIE, Y. and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. MR3099117 https://doi.org/10.1214/13-AOS1094

YAKIR, B. (1997). A note on optimal detection of a change in distribution. *Ann. Statist.* **25** 2117–2126. MR1474086 https://doi.org/10.1214/aos/1069362390